图书情报工作
第64卷 第24期 2020年12月

专题情报数据管理与智能分析平台的构建*

■ 于倩倩[1,2] 钱力[1,2] 程冰[1] 常志军[1,2] 王慧丽[3] 靳茜[4]

[1] 中国科学院文献情报中心 北京 100190 [2] 中国科学院大学经济与管理学院图书情报与档案管理系 北京 100190
[3] 中国化工信息中心 北京 100029 [4] 中国农业科学院农业信息研究所 北京 100081

摘 要：[目的/意义]面对多学科领域、多类型用户的专题情报服务需求，建立专题情报数据管理与智能分析平台。实现专题情报分析的流程化和智能化，同时对融入专家智慧的专题情报分析过程数据进行管理，丰富服务模式，提升服务需求响应速度。[方法/过程]在调研已有相关研究与实践分析基础上，提出平台设计思路、建设框架，对平台主要功能和关键技术进行剖析。[结果/结论]专题情报数据管理与智能分析平台已建设完成。平台集成了多来源多类型数据，打通了从数据到分析的服务链条。嵌入了多种情报分析方法和深度学习算法，实现了多维多层次分析服务。能够对分析过程和情报分析人员历史积累数据进行管理，实现数据共享和重复利用。

关键词：专题情报 数据管理 智能分析 情报分析
分类号：G250.2
DOI：10.13266/j.issn.0252-3116.2020.24.011

1 引言

在大数据情报分析和知识服务时代，专题情报服务正处于颠覆性变革阶段。多源异构数据的获取、人工智能技术的发展为专题情报分析带来新的契机[1]。国家科技图书文献中心(National Science and Technology Library, NSTL)面向国家战略需求提供专题情报服务已有十余载，积累了大量的专题情报分析过程数据，但这些数据一直处于分散自存储状态。如何将这些融入科技情报专家智慧的数据进行统一管理，为实现专题情报分析过程快速复现、专业信息共享和提供新型数据服务建立基础，是值得考虑的问题。NSTL、中国科学院文献情报中心(以下简称“文献情报中心”)通过多种方式获取了多源异构资源并进行了汇聚融合[2-3]。如何对已汇聚的科技大数据资源价值进行充分挖掘利用，弥补人工为主进行数据源遴选、数据采集、数据装载和数据分析的不足，建立基于多源数据计算的专题情报分析快速响应机制，是另一个值得考虑的问题。

本文在梳理分析现有相关平台软件、需求痛点、工作流程等的基础上，利用“专家+平台+数据”模式，以大数据技术、人工智能技术发展为契机，驱动建设专题情报数据管理与智能分析平台。一方面对NSTL情报分析人员线上或线下的高价值中间分析结果数据进行统一存储和管理；另一方面充分利用NSTL和文献情报中心建设的结构化和规范化数据，集成多种情报分析方法和深度学习算法，打通从数据到分析的服务链条。探索多种形式分析服务，实现在线专题情报分析服务的流程化和智能化。以期为相关情报研究提供平台工具抓手，为服务平台建设提供参考和借鉴。

2 相关研究与实践分析

大数据环境下，用户需求与应用场景越来越重要。本文对专题情报分析工具、数据管理工具及相关研究进行工具调研和文献调研，并与NSTL一线专题情报服务人员进行交流访谈，对用户需求进行深入挖掘，为

* 本文系NSTL资助项目“专题情报数据协同管理与分析服务”(项目编号:科1926)研究成果之一，受“中国科学院文献情报中心成立七十周年主题论坛与纪念文集出版”项目资助出版。

作者简介：于倩倩(ORCID：0000-0001-8777-1171)，馆员，博士研究生；钱力(ORCID：0000-0002-0931-2882)，研究馆员，博士，通讯作者，E-mail：qianl@mail.las.ac.cn；程冰(ORCID：0000-0002-0628-2585)，副研究馆员，博士；常志军(ORCID：0000-0001-9211-8599)，副研究馆员，博士研究生；王慧丽(ORCID：0000-0001-7276-4032)，工程师，硕士；靳茜(ORCID：0000-0002-6066-33653)，副研究馆员，硕士。

收稿日期：2020-11-05 修回日期：2020-12-14 本文起止页码：92-105 本文责任编辑：王传清

平台应用场景设计建立基础。

2.1 专题情报分析工具研究

专题情报分析工具按照能否在线提供服务,可分为平台类工具和软件类工具。平台类工具包括专门提供情报分析服务的平台(如 InCites[4]、SciVal[5]、Incopt[6]、wizdomAI[7]等)以及科技文献检索平台增加分析评价功能(如 Dimensions[8]、Web of Science[9]、CNKI[10]、万方数据知识服务平台[11]等),后一类平台服务方向从知识发现向知识评价过渡。软件类工具包括美国德雷塞尔大学的 CiteSpace[12]、荷兰莱顿大学的 VOSviewer[13]、印第安纳大学的 Sci2[14]、开源工具 Gephi[15]、科睿唯安的 DDA[16]、瑞典于默奥大学的 Bibexcel[17]等。

刘斐等[18]认为 InCites 平台、SciVal 平台包含了大量评价指标,能够承担大部分科研影响力分析评价工作。许景龙等[19]认为智能语义检索、集成和灵活的数据处理、综合化分析视角、内容智能化的自动报告是 Incopt 等专利情报分析工具的主要发展趋势。C. Herzog 等[20]指出 Dimensions 将多类型数据(出版物、专利、基金项目、政策、临床试验)和不同维度分析(趋势分析、研究人员分析、基金项目分析、机构分析、对比分析)集成在一个平台,期望集成促进创新。泰勒-弗朗西斯出版集团(Taylor & Francis Group)利用大数据分析与机器学习技术,研发 wizdomAI[7],涵盖出版物、专利、基金项目、机构、作者等多类型数据,为科研人员与研究机构提供面向全价值链的深度分析服务。于晓彤等[21]研究发现 CiteSpace、VOSviewer、DDA、Bibexcel 等在知识图谱研究中得到了高频应用。杨静等[22]研究发现 Sci2 适合大量数据的去重,网络输出可编辑能力强。邓君等[23]认为 Gephi 更适用于处理动态大数据,可视化功能强大。

从已有专题情报分析工具实践与相关研究来看,平台类工具向多源化、智能化、细粒度分析方向发展,多源异构数据汇聚融合成为情报分析的新型数据基础设施,利用人工智能技术手段挖掘知识成为新的增长点。软件类工具功能各有特色,在数据清洗、可视化分析等功能点上有很多值得借鉴的地方。但完成一个报告,从数据获取到数据分析,往往需要在多个软件工具间切换[24],通常无法实现一站式操作,也不具备分析过程数据管理功能。已有研究[25-26]指出,国外的情报分析工具较多,但部分产品存在价格高、出口限制或知识产权壁垒等问题。国内相关工具研发不足,在情报研究中发挥作用有限,研发投入有待提高。因此,建设具有自主知识产权的情报分析工具十分必要。

2.2 数据管理工具研究

在数据管理工具方面,最具显示度的是科学数据管理平台的研究与实践。国内外科学数据管理平台建设快速发展,包括哈佛大学 Dataverse[27]、Dryad 数据仓储[28]、澳大利亚国家数据服务网 ANDS[29]、中国科学院数据云[30]、北京大学开放研究数据平台[31]、武汉大学科学数据管理平台[32]等。

崔旭等[33]认为数据管理平台核心服务功能包括数据管理计划、数据创建、数据存储、数据获取、数据分析、数据共享,Dataverse、ANDS 具有上述所有功能。卫军朝等[34]认为国内的科学数据管理平台多是数据主导型平台,主要是对用户已经生成的科学数据进行存储和管理,如武汉大学科学数据管理平台、中国科学院数据云。朱玲等[35]比较发现 Dataverse、Dryad 均面向多学科,但前者以社会科学为主,元数据方案以 DDI 元数据标准为基础扩展而成;后者以生物科学、生态科学为主,元数据方案遵循 DC 元数据标准。

从已有研究和平台存储的数据来看,不同的数据管理平台功能特点不同,学科范围重点不同,元数据方案也有所差异。有些数据管理平台以本机构科学数据的管理和保存为目标,如国内部分数据管理平台。有些以收集和管理社会不同机构的科学数据为目标,如 Dryad、ANDS。目前,武汉大学科学数据管理平台中的计量分析研究数据集中存储了 5 条情报分析相关数据,数据描述字段包括题名、作者、日期、相关描述、URI、所属数据集,附件为统计分析数据集、分析报告等。北京大学开放研究数据平台中也有情报分析相关数据,但存储比较分散,隶属于不同的数据空间和数据集。数据描述字段包括题名、作者、联系人、提交者、提交日期、描述、学科等,主要是对分析数据集的存储。总体而言,情报分析过程数据逐渐得到重视,但重视程度还远远不够。

2.3 专题情报服务实践分析

笔者分别与来自 NSTL 成员单位(中国科学院文献情报中心、中国科学技术信息研究所、中国医学科学院医学信息研究所、中国化工信息中心等)一线从事专题情报服务的 8 位情报分析人员进行了交流访谈,主要了解当前专题情报服务中的痛点及情报分析人员的需求。调研发现,在专题知识组织体系构建、数据获取、数据清洗、数据分析、数据管理等方面均存在短板和需求痛点。

在专题知识组织体系构建方面,基本靠情报分析

人员手工完成资料的收集与整理,依赖专家指导形成体系,缺少自动化辅助工具。在数据获取方面,多元数据的获取(基金项目、政策数据等非传统文献)成为趋势,但非传统文献数据源分散,需人工去不同网站检索收集。数据批量获取困难,如 WoS 单次下载限制在 500 条,每个查询最多下载 10 万条[36]。在数据量大时,下载时长及人力消耗大。在数据清洗方面,尽管相关工具起到了一定的辅助作用,但处理能力有限。清洗方式主要依赖规范词表和规则,而情报分析人员累积的词表处于自存储、自管理、自使用状态。在数据分析方面,目前的分析工具在大数据量分析时存在困难,通常超过 5 万条数据时,工具运算速度慢。超过 10 万条数据时,容易卡机[37]。在数据管理方面,专题情报分析的过程文件通常留存在课题组或个人手中,缺乏数据管理规范和数据管理平台,难以实现数据共享。

3 平台设计与实现

3.1 平台设计思路

专题情报研究是针对特定用户特定需求的情报研究工作[38]。由于情报问题与任务往往具有很强的动态性与个性化,这个特点导致很难生产出一套通用的情报分析系统[39]。笔者在借鉴现有相关研究和实践分析基础上,提出专题情报数据管理与智能分析平台的设计思路:①将期刊论文、会议论文、专利、基金项目等多来源、多类型数据集成为一体,充分利用已汇聚的科技大数据资源,支持多元数据的获取;②建立人机结合的数据获取与数据清洗途径,借助相关工具和算法,辅助情报分析人员建立知识组织体系、检索式以及自动化数据清洗;③利用大数据技术,提升大数据量数据分析的速度;④设计多维多层次分析模式,集成多种情报分析方法和深度学习算法,智能化生成导出报告;⑤对线上或线下的高价值中间分析结果数据进行统一存储和管理,实现专题情报分析数据的平台化管理、重复性利用以及知识的可积累。

3.2 平台整体架构

根据平台设计思路,确定专题情报数据管理与智能分析平台的整体架构(见图 1)。该架构包括大数据基础架构、大数据资源体系、专题数据获取与清洗规范、专题情报分析计算模型、数据管理与分析服务 5 个层次。

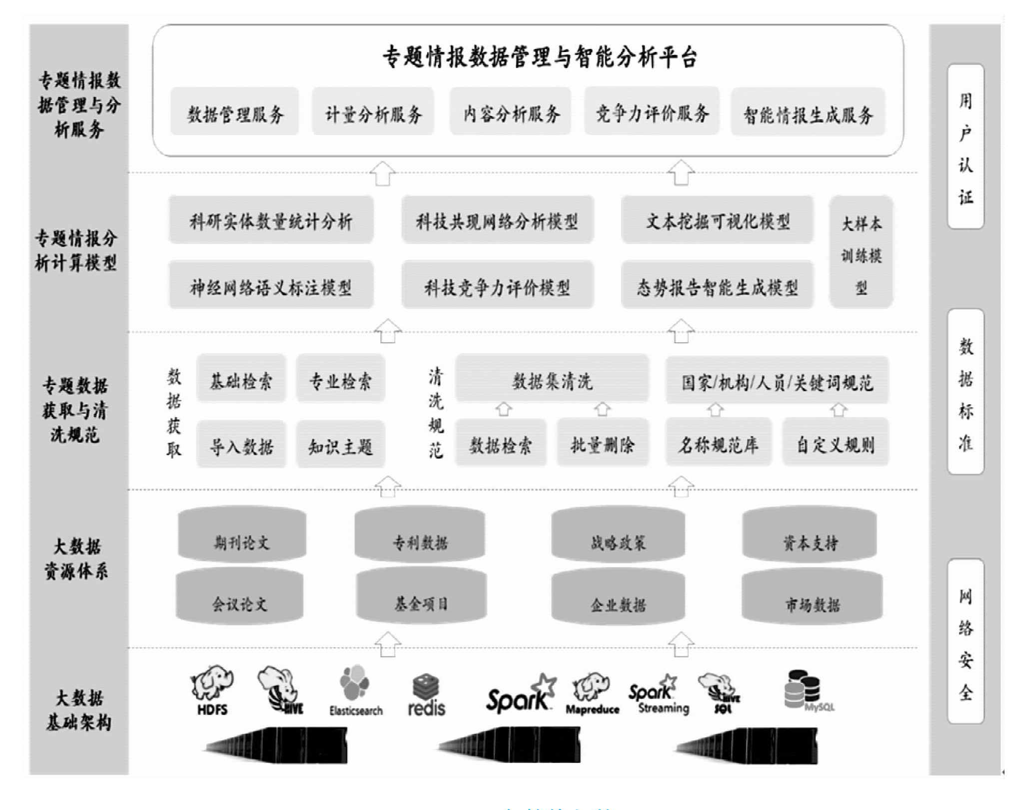


图 1 平台整体架构

3.2.1 大数据基础架构

中国科学院文献情报中心基于开源 Apache Hadoop 生态群技术,建设了科技大数据基础平台,对海量科技资源进行汇聚融合。笔者将科技大数据基础平台的多源汇聚融合资源,作为专题情报数据管理与智能分析平台的基础数据来源。Elasticsearch 是基于 Lu-

cene 的分布式、可扩展、高实时的搜索与数据分析引擎,在专题情报数据管理与智能分析平台中用于存储基础数据以及存储专题检索、导入的数据,并支撑检索结果展示及情报分析。Redis 是一个高性能的 key-value 数据库,用于缓存用户访问数据,并为平台性能优化提供支持。MySQL 用于专题情报分析过程数据的保存。

3.2.2 大数据资源体系

平台拥有"期刊论文＋会议论文＋专利数据＋基金项目＋政策数据＋企业数据＋资本支持＋市场数据"等多来源多类型的数据资源体系。其中,期刊论文、会议论文、专利数据为 NSTL 及文献情报中心与国内外出版商、相关信息机构等第三方协商获取、交换、购买等方式建设的数据资源。期刊论文、会议论文数据体量达 1.1 亿多条,专利数据体量达 8 000 多万条。基金项目数据为自采集的 10 余个国家的基金项目,包括美国国家自然科学基金(NSF)、中国国家自然科学基金(NSFC)等,共计 520 多万条。政策数据为存缴的国内相关政策信息,共计 26 万余条。企业数据、资本支持、市场数据提供数据存储支持,目前所获得的数据量相对较少。

3.2.3 专题数据获取与清洗规范

平台建立了人机结合的数据获取与清洗规范途径,面向专题研究领域、方向等,辅助用户梳理与构建权威、全面的数据资源。平台能够支持多类型数据的统一描述表示与存储管理,对用户通过检索、导入、知识主题筛选等方式获取的专题数据进行解析、集成汇聚、排重融合、清洗规范。用户在平台自动化处理数据基础上,能够对获取到的数据集以及科研实体如国家/地区、机构、人员、关键词等进行人工编辑处理及设置规则,优化平台自动化处理效果。

3.2.4 专题情报分析计算模型

平台嵌入了多种算法模型,包括科研实体统计分析模型、共现网络分析模型、文本挖掘可视化模型、神经网络语义标注模型、大样本训练模型、科技竞争力评价模型等。通过数据、算法和计算驱动智能化分析,从而实现专题情报数据智能计算＋情报专家智慧结合的情报分析报告的快速生产与递送。

3.2.5 数据管理与分析服务

平台提供专题情报分析过程数据及本地数据的管理功能,提供计量分析、内容分析、竞争力评价分析等多种形式分析服务。不同的分析服务对应的数据资源类型不同,采用的分析计算模型也有所不同。平台支持计量分析、内容分析维度的筛选,支持分析报告的智能生成导出。支持计量分析、内容分析和竞争力评价分析维度的前端页面发布和展示。

3.3 主要功能与关键技术

专题情报数据管理与智能分析平台包括专业情报分析、快速分析、竞争力评价分析、数据管理 4 个主要的功能模块。快速分析借鉴 Web of Science、CNKI 等在文献检索基础上增加计量分析功能,面向通过平台审核的用户提供服务,供用户快速了解领域概况。专业情报分析、数据管理面向情报分析人员提供服务。竞争力评价分析面向特定情报分析人员提供服务。

3.3.1 专业情报分析

根据对情报分析人员的分析流程调研,笔者将平台专业情报分析功能分为创建专题－专题知识组织－专题数据汇聚－专题数据清洗规范－专题情报分析 5 个步骤,如图 2 所示:



图 2 专业情报分析流程

(1)创建专题。在创建专题步骤,用户能够浏览已经创建的专题列表,可以查看已经创建的专题名称、数据量、数据时间范围、数据类型、专题创建时间、状态等,也可以根据需要创建新的专题。如图 3 所示:

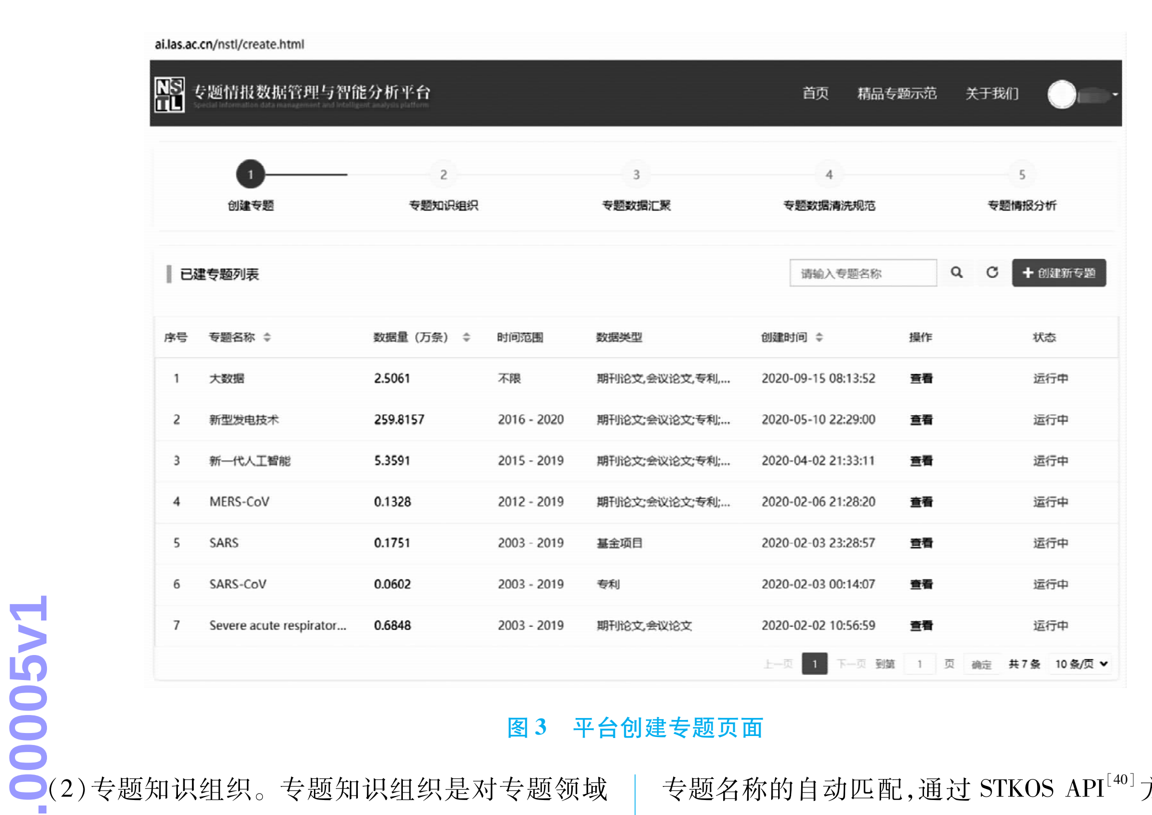


图 3　平台创建专题页面

（2）专题知识组织。专题知识组织是对专题领域的研究范畴和知识体系进行组织和管理，为下一步专题数据汇聚提供知识主题。如何通过自动化方式辅助构建知识组织体系和检索式，提高情报分析人员的工作效率，是平台建设需要考虑的一个问题。

在专题知识组织步骤，平台利用 STKOS 词表实现专题名称的自动匹配，通过 STKOS API[40] 方式获取上下位类、同义词，辅助建立知识组织体系。支持用户对平台自动推荐的知识组织体系概念、标签词进行编辑、修改、删除。同时支持导入或节点添加方式建设知识组织体系。如图 4 所示：

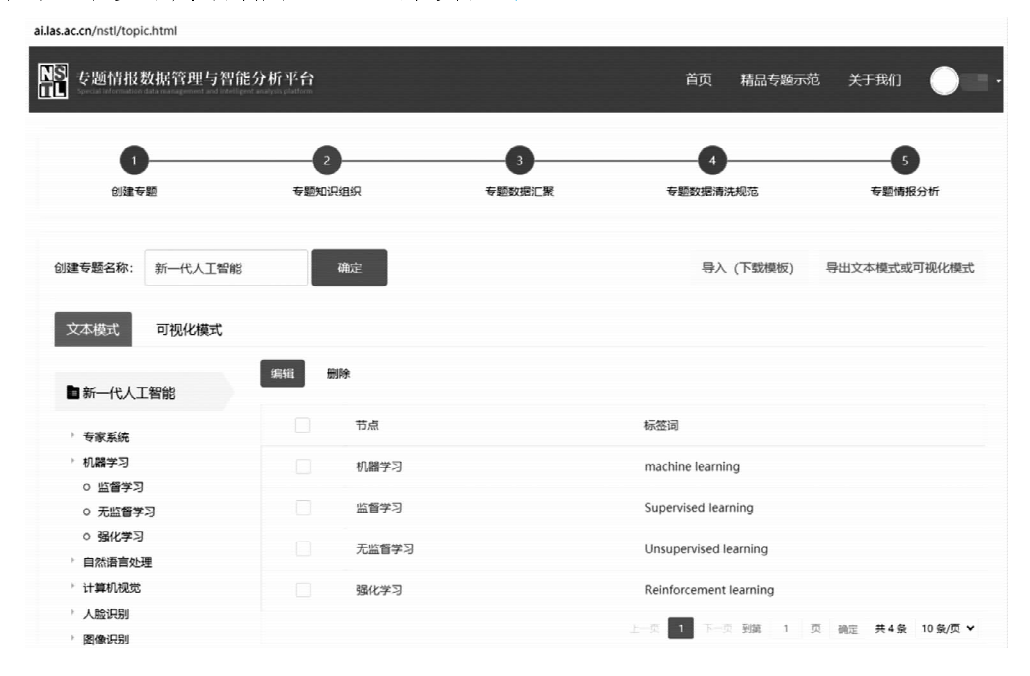


图 4　平台专题知识组织页面

（3）专题数据汇聚。在专题数据汇聚步骤，平台将期刊论文、会议论文、专利、基金项目等多来源多类型数据集成为一体。平台支持对多种类型数据的并行检索、清洗规范和情报分析。也就是说用户通过一套流程化操作，可以得到不同类型数据基础的分析报告，以提升专题情报分析速度和效率。

平台支持用户通过基础检索发现、专业检索式或本地导入方式获取数据,支持通过选择知识主题(专题知识组织体系中的节点,默认选择节点、同义词及下位类)自动生成检索式。如图5所示。检索结果按照相关性、时间进行排序,从多角度对获取的数据进行分面揭示。



图5　平台专题数据检索页面

(4)专题数据清洗规范。数据清洗规范是情报分析工作的重要步骤,是保证分析结果准确可靠的前提条件。在专题数据清洗规范步骤,平台通过去重、检索、排序、删除等方式对数据集进行清洗规范。平台允许用户遵循合理使用的原则,导出数据,单次可导出5 000条。在科研实体自动清洗规范方面,平台以全量数据自动清洗规范结果为基础,应用到所检索获取的数据集。相较于只针对数据集的清洗规范,更能挖掘科研实体之间的关联,提升规范效果。

平台以世界各国和地区名称规范代码表为基础,对国家/地区进行自动清洗规范。采用层次化混合结构的深度学习框架模型,利用单层双向LSTM网络向量语义匹配、字符编辑距离结合的方式,对机构名称相似度进行计算,辅以国家、城市、邮编、机构名称排序特征,对机构进行自动清洗规范;采用作者名称消歧规则集合,对作者进行自动清洗规范[41]。借助STKOS词表的规范概念和同义词,对关键词进行自动清洗规范。平台支持对非规范名称的人工编辑规范以及对多个非规范名称合并的功能,默认按照发文量对规范名称进行排序。见图6。

(5)专题情报分析。在专题情报分析步骤,考虑到大数据时代的科技情报工作,单一维度的信息分析难以满足需求,需要以多维度的视角从数据和方法上实现创新[42],平台设置了计量分析模块和内容分析模块。不同的数据类型对应不同的分析维度,嵌入Echarts、Gephi等开源工具对分析结果进行可视化,显示方式为折线图、柱状图、气泡图、堆积图、网络图、词云图、地图等,分析结果图可下载。

在计量分析模块,主要是对年代、国家/地区、机构、作者、关键词、技术构成等结构化内容进行统计分析、合作网络分析、共词分析,如图7所示。论文、专利、基金项目分别具有15种、21种、13种分析维度。在内容分析模块,主要是从非结构化科技文献内容中,采用主动学习指导的深度学习抽取框架对研究问题、关键技术以及相互之间的关系进行抽取[43],从语义层面丰富专题情报智能分析体系。对抽取的研究问题、关键技术进行数量统计分析和关联分析,如图8所示。论文、专利、基金项目分别具有11种、16种、11种分析维度。

用户可对不同类型数据的分析维度进行筛选,自动生成和导出情报分析报告。分析报告上带有NSTL图标等产权特征,报告内容包括分析检索式、数据量、

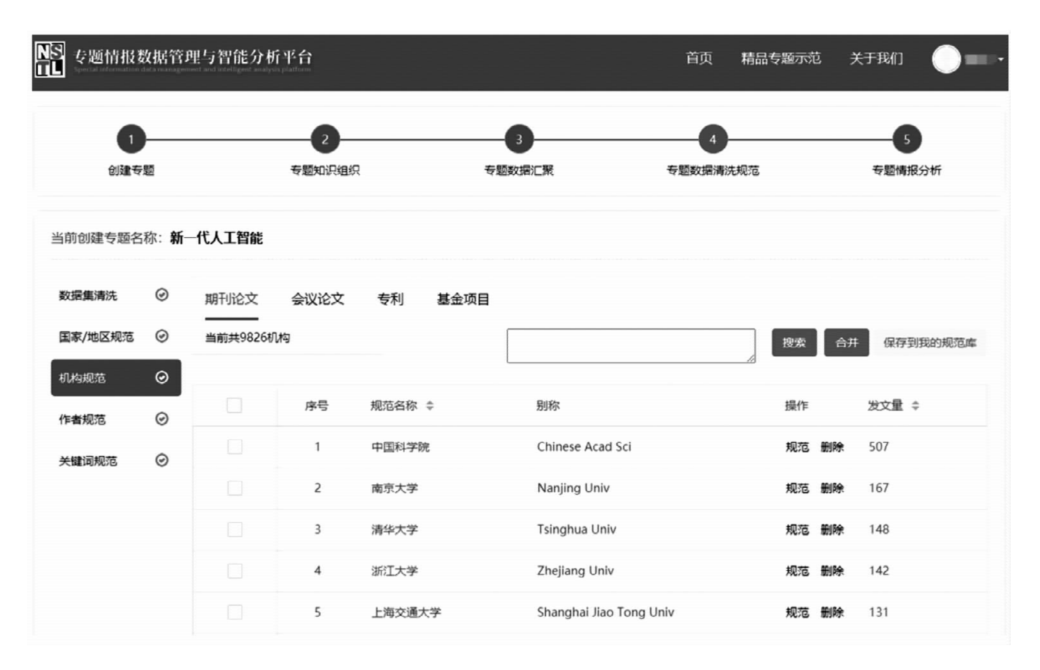


图 6　平台专题数据清洗规范页面

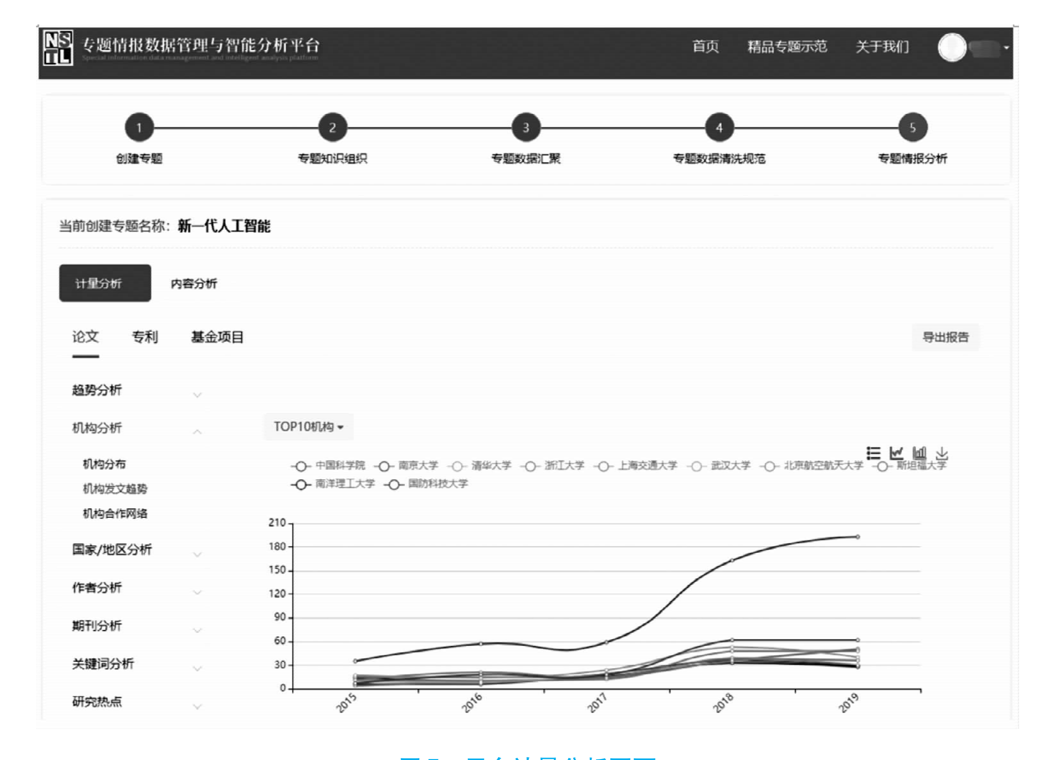


图 7　平台计量分析页面

文献类型、时间范围，以及计量分析或内容分析的分析图表结果数据和相关文字内容描述。用户还可选择要公开的分析维度和专题，在平台前端发布，供其他用户浏览专题的计量分析和内容分析结果，较为深入地了解专题发展态势。

3.3.2　数据管理

根据与情报分析人员的交流访谈以及对现有情报分析数据的存储现状分析，笔者将分析检索式、数据集、规范数据、分析报告纳入专题情报数据管理与智能分析平台的数据管理范畴。分析检索式为情报分析人员检索数据时所用到的检索式，数据集为检索结果数据集或经人工参与处理后的数据集，规范数据包括国家/地区规范数据、机构规范数据、作者规范数据、关键词规范数据，分析报告为系统自动生成报告或经情报分析人员加工撰写的报告。

笔者设计了在线专题情报分析与专题情报数据管



图 8　平台内容分析页面

理的交互方式，满足专题情报分析过程中的高价值中间分析结果数据的实时保存和管理。用户点击保存检索式按钮，能够将专题数据汇聚步骤的检索式自动保存到检索式管理列表。点击确定导入专题库按钮，能够将检索结果数据集自动保存到数据集管理列表。点击保存到我的规范库按钮，能够将专题数据清洗规范步骤的科研实体规范数据自动保存到规范数据管理列表。点击保存报告，能够将专题情报分析步骤自动生成的报告保存到分析报告管理列表。同时支持用户对本地数据进行上传导入，解决融入科技情报专家智慧数据处于分散自存储的问题。如图 9 所示：

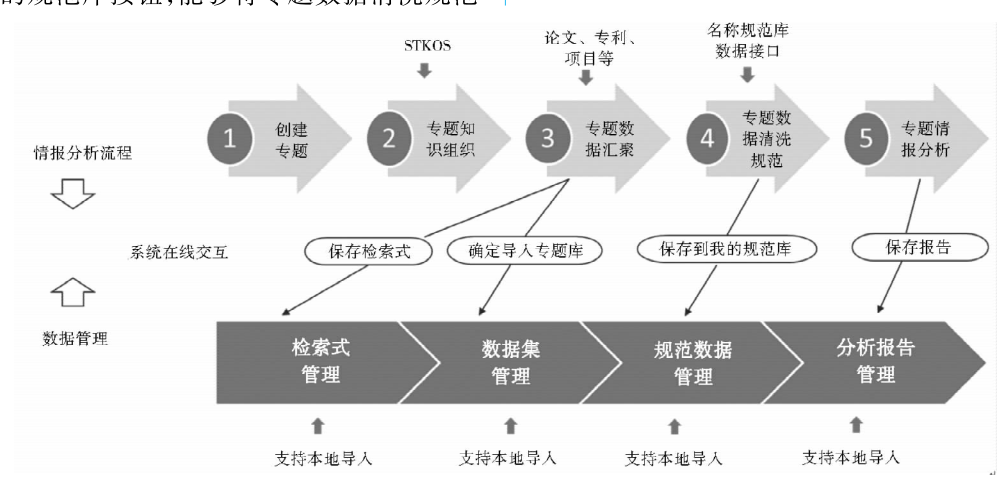


图 9　数据管理与专业情报分析流程交互

用户在专题情报分析过程中实时保存的规范数据或本地上传的规范数据，将自动应用于后续所建专题的科研实体数据清洗规范，辅助提升后续专题科研实体数据清洗规范效果。不同的数据类型描述方式有所不同，分析检索式描述字段包括检索式、检索词、所属专题、标签、创建时间，规范数据描述字段包括规范名称、其他名称、创建时间，数据集描述字段包括数据集体量、获取方式、所属专题、数据年限、数据类型、创建时间，分析报告描述字段包括报告名称、报告类型、数据年限、所属专题、生成方式、创建时间。如图 10 所示：



图 10　数据集管理

平台支持用户对所管理的分析检索式、数据集、规范数据、分析报告进行导出，实现数据的重复利用。平台设置了数据共享规则，数据默认为不公开状态，若用户点击了公开按钮，则会将数据共享给平台其他用户。在选择公开后，若由于某些原因不想或不便公开，可再选择不公开，平台会将已公开数据撤回，数据重新回到不公开状态。点击数据集管理中的数据体量(万条)数字，则可实现专题情报分析过程的快速复现。

3.3.3　竞争力评价分析

为丰富专题情报分析形式，拓展分析服务层次，平台增加了竞争力评价分析功能，能够对不同国家或地区的专题领域发展水平进行分析评价。竞争力评价分析功能遵循数据来源广泛、评价内容全面、有总体分析和分项分析、多样化的指标体系、多种可视化表现方式等评价原则和方法[44]。

竞争力评价模型是竞争力评价分析的核心，笔者设计了 5 个一级指标、13 个二级指标、14 个三级指标的竞争力评价指标体系，如图 11 所示。平台支持评价模型管理和对应的指标体系管理，能够对评价模型进行增、删、改，能够按照评价模型名称、类型等对模型进行检索，能够对指标进行增、删、改、权重设置等。见图 12。

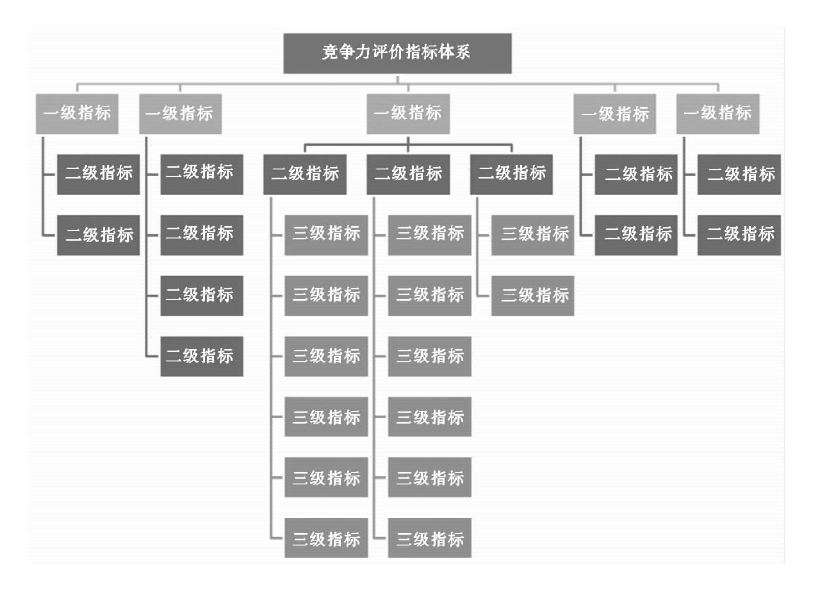


图 11　竞争力评价指标体系示意

于倩倩，钱力，程冰，等. 专题情报数据管理与智能分析平台的构建[J]. 图书情报工作,2020,64(24):92－105.

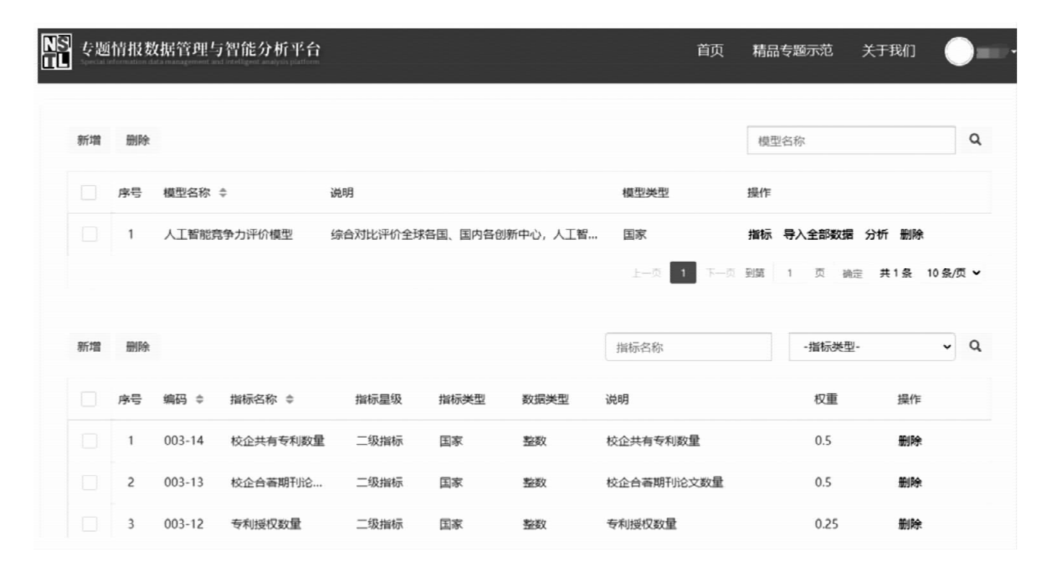


图12　评价模型管理

由于专题国家或区域竞争力评价分析往往会涉及到产业数据，而此部分数据并非科技大数据基础平台的优势所在，专题情报数据管理与智能分析平台重点对专题国家或区域竞争力分析维度、综合评价结果进行可视化展示和图表揭示。笔者根据所设计的竞争力评价指标体系，设置了竞争力评价分析所需数据的组织方式模板，供平台用户下载使用。用户根据该模板组织数据，并导入上传到平台中，则可进行相应的数据可视化展示。

竞争力评价分析结果可在平台前端发布，供其他用户了解不同国家或地区的专题领域发展水平。可视化图表可下载，辅助提升情报分析人员撰写或生产专题竞争力分析报告的速度和效率。

3.4　实现效果

专题情报数据管理与智能分析平台是 NSTL2019 年部署的牵引新型数据服务与情报分析服务于一体化的信息化示范平台(http://ai. las. ac. cn)。平台基于 B/S 架构，使用 Java 语言开发，采用基于 Springcloud + Springboot 的 Web 微服务框架，数据存储使用 Elasticsearch、MySQL、Redis 等数据库混合存储框架，展示采用可组件化的 Vue 进行页面交互。目前平台已经完成一期研发，并对外发布试运营。

3.4.1　充分利用已汇聚科技大数据资源，一套操作流程可获取多类型分析报告

平台突破了单一数据源或单一数据类型的限制，充分利用已汇聚的科技大数据资源，将期刊论文、会议论文、专利、基金项目等多来源、多类型数据集成为一体。平台界面友好、操作简单，设计了向导式的专业情报分析过程步骤，实现专业情报分析的流程化管理。用户通过一套流程化操作，可获得不同类型数据基础的分析报告。相较于科技文献检索平台，更多地嵌入了数据清洗规范功能和不同类型数据的管理功能。相较于情报分析软件，实现了从数据检索到数据分析的整个流程，既支持本地检索，也支持数据导入，弥补了情报分析软件仅支持数据导入的不足。

新冠疫情爆发期间，笔者通过平台 Demo 版本的专业情报分析功能，快速分析与生成了以论文、专利、项目为基础的 MERS-Cov 数据分析报告和 SARS-Cov 数据分析报告，引起了业界的关注和共鸣。负责“先进轨道交通”专题的情报分析人员利用平台的全流程自动化情报研究报告生产机制，完成了相关分析报告，为轨道交通行业相关用户提供了更快速、精准、全面的情报支撑服务[45]。在平台前端发布的专题(MERS-Cov、SARS-Cov)如图13所示，专题名称右侧显示分析所用的数据类型，下侧显示部分分析结果。点击右侧更多按钮，用户登录后可以看到该专题的所有计量分析和内容分析结果。

3.4.2　实现了多维多层次分析，集成了多种情报分析方法和深度学习算法

平台支持多维多层次分析模式，既包括计量分析、内容分析，也包括竞争力评价分析，满足不同类型的专题情报分析需求。计量分析利用统计分析、合作网络分析、共词分析等情报分析方法，对发文趋势、科研合作网络及研究热点等进行揭示。专题数据清洗规范、内容分析分别利用不同的深度学习算法对科研实体消

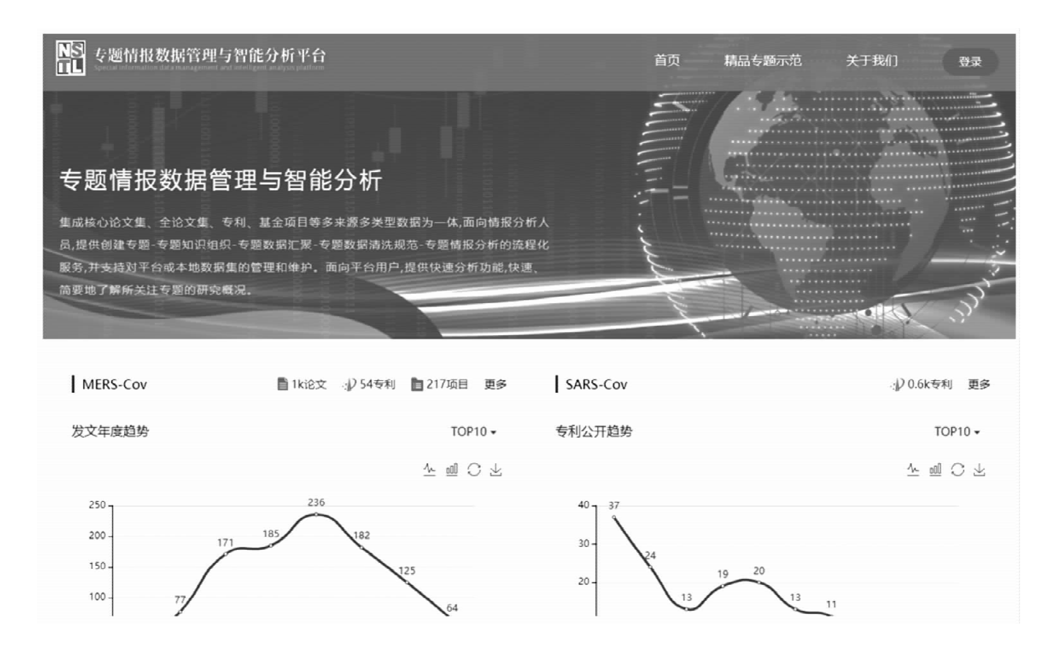


图 13　平台首页

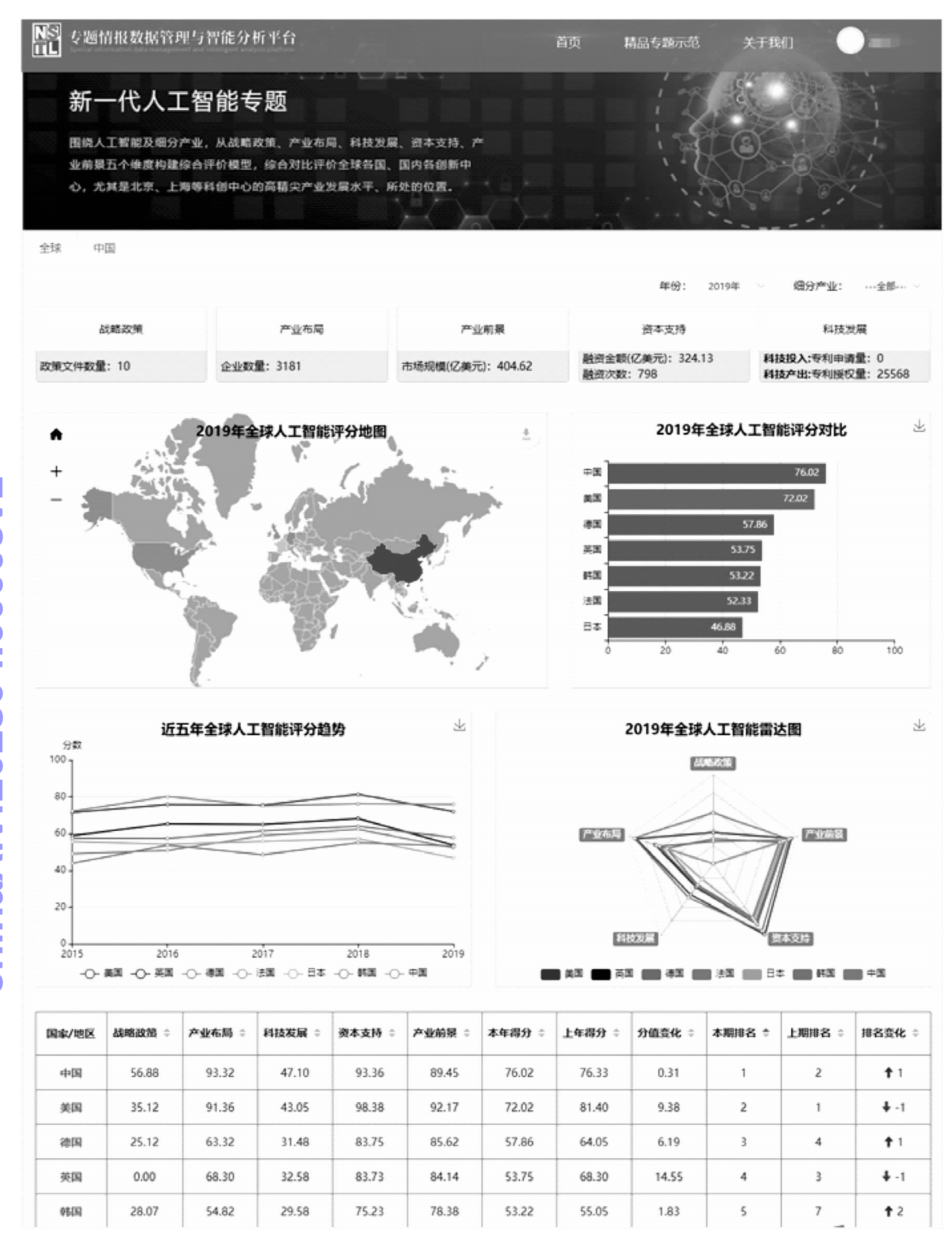
歧归一、对非结构化文献内容中的研究问题和关键技术进行识别抽取。竞争力评价分析利用统计分析、对比分析、综合评价等方法对指标分项结果和综合分析结果进行展示。

以"新一代人工智能"竞争力评价分析应用示范实现为例,平台围绕新一代人工智能及细分产业,以战略政策、产业布局、科技发展、资本支持、产业前景为一级指标,并相应建立了 13 个二级指标、14 个三级指标,构建了竞争力评价模型,综合对比评价全球各国、国内各创新中心的产业发展水平。全球各国新一代人工智能产业发展综合分析评价结果如图 14 所示。其中战略政策分析评价政策支持力度及趋势,所用数据为政策类数据。产业布局分析评价基础层、技术层、应用层等企业分布,所用数据为企业数据。科技发展包括科技投入、产出水平以及科技合作水平,所用数据为论文、专利、项目数据。资本支持评价社会资本投入情况,所用数据为投融资类数据。产业前景评价分析产业市场发展潜力,所用数据为市场数据。点击战略政策、产业布局、科技发展等一级指标,可以看到二级指标或三级指标形成的分析维度。

此外,平台明确了数据管理对象,能够对专题情报分析过程中产生的数据以及用户本地数据进行保存、管理。目前,平台处于推广试运营阶段,用户 40 余位,来自 18 个单位(NSTL 成员单位或服务站)。平台在数据精准度、数据计算速度以及用户体验等方面仍然有比较大的改善空间。项目组将以"边服务、边建设、边完善"的组合方式,对平台功能进行优化和完善,为 NSTL 专题情报服务提供新的支撑和发力点。

4　结语

数据服务与平台工具是未来智能情报模式转型升级的必须阶段,更是现阶段有效解决情报分析人员面临体量大、需求类型多、任务要求紧急的情报服务需求的新路径与新方法。专题情报数据管理与智能分析平台是具有自主知识产权的数据管理与情报分析工具,将多来源多类型数据集成进来,为深度挖掘和释放多源汇聚的科技数据资源价值提供抓手。提供多维多层次分析服务,将情报分析过程流程化和智能化。打通从数据到分析的服务链条,探索多样化分析服务实现方式,无缝嵌入多种情报分析方法和深度学习算法,同时对分析过程数据和情报分析人员的历史积累数据进行管理,丰富服务模式,提升情报分析人员服务能力,提高服务需求响应速度。未来,将从文本分析、语义分析等角度探索更多复杂情报分析需求的解决方案,对平台进行持续优化、完善和迭代升级,将平台建设成为情报分析人员的常用工具,帮助情报分析人员更好更快速地完成情报分析工作。



国家/地区	战略政策	产业布局	科技发展	资本支持	产业前景	本年得分	上年得分	分值变化	本期排名	上期排名	排名变化
中国	56.88	93.32	47.10	93.36	89.45	76.02	76.33	0.31	1	2	↑1
美国	35.12	91.36	43.05	98.38	92.17	72.02	81.40	9.38	2	1	↓-1
德国	25.12	63.32	31.48	83.75	85.62	57.86	64.05	6.19	3	4	↑1
英国	0.00	68.30	32.58	83.73	84.14	53.75	68.30	14.55	4	3	↓-1
韩国	28.07	54.82	29.58	75.23	78.38	53.22	55.05	1.83	5	7	↑2

图 14　新一代人工智能竞争力评价分析结果

参考文献：

[1] 刘细文,王丽. 面向国家重点科技项目的专题情报服务十年探索[C]//国家科技图书文献中心成立二十周年文集. 北京:科学技术文献出版社, 2020:346－349.

[2] 鲜国建,罗婷婷,赵瑞雪,等. 从人工密集型到计算密集型:NSTL 数据库建设模式转型之路[J]. 数字图书馆论坛,2020(7):52－59.

[3] 钱力,谢靖,常志军,等. 基于科技大数据的智能知识服务体系研究设计[J]. 数据分析与知识发现, 2019, 3(1):4－14.

[4] InCites[EB/OL]. [2020－09－22]. https://incites. clarivate. com.

[5] SciVal[EB/OL]. [2020－09－22]. https://www. scival. com/.

[6] Incopt[EB/OL]. [2020－09－22]. https://www. incopat. com/.

[7] wizdom. ai[EB/OL]. [2020－09－22]. https://www. wizdom. ai/.

[8] From idea to impact-The next evolution in linked scholarly information [EB/OL]. [2020-09-22]. https://www.dimensions.ai/.
[9] Web of Science[EB/OL]. [2020-09-22]. http://apps.webofknowledge.com.
[10] CNKI[EB/OL]. [2020-09-22]. https://www.cnki.net/.
[11] 万方数据知识服务平台[EB/OL]. [2020-08-03]. http://www.wanfangdata.com.cn/index.html.
[12] Citespace: visualizing patterns and trends in scientific literature [EB/OL]. [2020-10-20]. http://cluster.ischool.drexel.edu/~cchen/citespace/download/.
[13] Vosviewer visualizing scientific landscapes[EB/OL]. [2020-10-20]. https://www.vosviewer.com/.
[14] Sci2 tool[EB/OL]. [2020-10-20]. https://sci2.cns.iu.edu/user/index.php.
[15] The open graph viz platform[EB/OL]. [2020-10-20]. https://gephi.org/.
[16] Derwent data analyzer[EB/OL]. [2020-10-20]. https://clarivate. com/derwent/solutions/derwent-data-analyzer-automated-ip-intelligence/.
[17] PERSSON O, DANELL R, SCHNEIDER J. How to use bibexcel for various types of bibliometric analysis[C]//Celebrating scholarly communication studies: a festschrift for Olle Persson at his 60th birthday. Leuven: International Society for Scientometrics and Informetrics, 2009:19-24.
[18] 刘斐,张美琦. InCites 平台及 SciVal 在科研影响力评价应用中的比较研究[J]. 图书馆杂志,2019,38(7):60-68.
[19] 许景龙,吕璐成,赵亚娟. 面向专利分析流程的专利情报分析工具功能比较研究[J]. 情报理论与实践,2020,43(8):178-185,151.
[20] HERZOG C, HOOK D, KONKIEL S. Dimensions: bringing down barriers between scientometricians and data[J]. Quantitative science studies,2020,1(1):387-395.
[21] 于晓彤,潘雪莲,华薇娜. 知识图谱研究中的软件引用和扩散分析[J]. 情报资料工作,2019,40(2):19-29.
[22] 杨静,程昌秀. 文献"大数据"分析软件 Citespace 和 Sci2 的对比分析研究[J]. 计算机科学与应用, 2017, 7(6):580-589
[23] 邓君,马晓君,毕强. 社会网络分析工具 Ucinet 和 Gephi 的比较研究[J]. 情报理论与实践, 2014, 37(8):133-138.
[24] 王丽. 开源/免费工具比较及专利分析全流程解决方案研究[J]. 情报理论与实践,2016,39(1):118-122.
[25] 刘玉琴, 汪雪锋, 雷孝平. 科研关系构建与可视化系统设计与实现[J]. 图书情报工作, 2015,59(8):103-110.
[26] 崔明,潘雪莲,华薇娜. 我国图书情报领域的软件使用和引用研究[J]. 中国图书馆学报,2018,44(3):66-78.
[27] Harvard Dataverse[EB/OL]. [2020-12-10]. https://dataverse.harvard.edu/.
[28] Dryad[EB/OL]. [2020-12-10]. https://datadryad.org/stash.
[29] Australian national data service[EB/OL]. [2020-12-10]. https://www.ands.org.au/working-with-data.
[30] 中国科学院数据云[EB/OL]. [2020-12-10]. http://www.csdb.cn/.
[31] 北京大学开放研究数据平台[EB/OL]. [2020-12-10]. https://opendata.pku.edu.cn/.
[32] 武汉大学科学数据管理平台[EB/OL]. [2020-12-10]. http://sdm.lib.whu.edu.cn/jspui/.
[33] 崔旭,赵希梅,王铮,等. 我国科学数据管理平台建设成就、缺失、对策及趋势分析——基于国内外比较视角[J]. 图书情报工作,2019,63(9):21-30.
[34] 卫军朝,张春芳. 国内外科学数据管理平台比较研究[J]. 图书情报知识, 2017(5):97-107.
[35] 朱玲,聂华,崔海媛,等. 北京大学开放研究数据平台建设:探索与实践[J]. 图书情报工作, 2016, 60(4): 44-51.
[36] MORALMUNOZ J A, HERRERAVIEDMA E, SANTISTEBANESPEJO A, et al. Software tools for conducting bibliometric analysis in science: an up-to-date review[J]. El profesional de la información, 2020, 29(1):1-20.
[37] 周超峰. 文献计量常用软件比较研究[D]. 武汉:华中师范大学,2017.
[38] 许军林,梁光德,钟红英,等. 高校图书馆专题情报产品生产质量控制研究[J]. 情报理论与实践,2013,36(6):68-72.
[39] 化柏林,李广建. 智能情报分析系统的架构设计与关键技术研究[J]. 图书与情报,2017(6):74-83.
[40] 王颖,张智雄,李传席,等. 科技知识组织体系开放引擎系统的设计与实现[J]. 现代图书情报技术,2015(10):95-101.
[41] 张建勇,钱力,于倩倩,等. 科研实体名称规范的研究与实践[J]. 数据分析与知识发现,2019,3(1):27-37.
[42] 滕广青,叶心,郭思月,等. 科技信息分析从单一维度到多维复合的演进[J]. 数字图书馆论坛,2019,12(12):2-8.
[43] 陶玥,余丽,张润杰. 科技文献中短语级主题抽取的主动学习方法研究[J]. 数据分析与知识发现, 2020, 4(10): 134-143.
[44] 李峰. 图书馆如何开展学科竞争力评价——由《英国科研表现之国际比较》报告得到的启示[J]. 大学图书馆学报,2015,33(2):72-76.
[45] 孙玉玲,秦阿宁,彭皓,等. 面向国民经济主战场-先进轨道交通技术情报监测与研究服务体会[C]//国家科技图书文献中心成立二十周年文集. 北京:科学技术文献出版社, 2020:449-451.

作者贡献说明:

于倩倩:平台需求调研、研究方案和功能体系设计,起草、撰写和修改论文;

钱力:提出平台建设框架和思路,提出论文修改意见,论文定稿;

程冰:对专题情报服务需求进行调研,参与分析场景设计;

常志军:平台多来源多类型数据维护和导入,平台技术选型;

王慧丽:参与新一代人工智能竞争力评价分析应用示范;

靳茜:提出平台建设目标、总体需求和平台完善思路。

于倩倩，钱力，程冰，等. 专题情报数据管理与智能分析平台的构建[J]. 图书情报工作,2020,64(24):92－105.

Research on the Construction of Data Management and Intelligence Analysis Platform for Subject Information

Yu Qianqian[1,2] Qian Li[1,2] Cheng Bing[1] Chang Zhijun[1,2] Wang Huili[3] Jin Qian[4]
[1] National Science Library, Chinese Academy of Sciences, Beijing 100190
[2] Department of Library, Information and Archives Management, School of Economics and Management, University of Chinese Academy of Sciences, Beijing 100190
[3] China National Chemical Information Centre, Beijing 100029
[4] Agricultural Information Institute of Chinese Academy of Agricultural Sciences, Beijing 100081

Abstract: [**Purpose/significance**] In order to meet the subject information service needs of multi-disciplinary and multi-type users, we construct a data management and intelligent analysis platform. Creating an intelligent analysis flow and managing different types of subject data that reflect experts' wisdom, the platform aims to enrich the service model and improve the service speed. [**Method/process**] On the basis of investigating existing relevant research and practice, we proposed the design idea and construction framework of the platform, and analyzed the main functions and key technologies. [**Result/conclusion**] The platform has been completed. It integrates multiple sources and multiple types of data, opens up the service chain from data to analysis, embeds a variety of information analysis methods and deep learning algorithms, realizes multi-dimensional analysis services. It can manage analysis process data and history data from information analysts, and then realize data sharing and reuse.

Keywords: subject information　data management　intelligent analysis　information analysis